Supporting Information

Morley *et al.* 10.1073/pnas.0808382106

SI Materials and Methods

Chromatin Immunoprecipitation (ChIP). Chromatin immunoprecipitation on 75–85% epiboly zebrafish embryos (AB strain) was performed as previously described (1) using anti-Ntl antibody or normal rabbit serum (NRS). Briefly, for each immunoprecipitation, embryos were enzymatically dechorionated and then fixed in 1.85% formaldehyde in 1X embryo medium for 20 min at room temperature. Glycine (0.125 M) was added to quench the formaldehyde and the embryos were washed in ice cold 1X PBS and snap frozen on liquid nitrogen or used immediately. Fixed embryos were homogenized in lysis buffer and incubated for 20 min on ice. Nuclei were collected by centrifugation, resuspended in nuclei lysis buffer, then incubated for 10 min before diluting with IP buffer and sonicating the chromatin sample on an ice bath. Sonication conditions were optimized to give fragments of ≈300–700 bp. The lysate was incubated overnight at 4 °C with 100 μL of protein G magnetic beads (Dynal) that had been prebound to the antibody. Beads were washed 5 times with RIPA buffer and once with 1X TBS at 4 °C. Bound complexes were eluted from the beads at 65 °C with vortexing in elution buffer. Cross links were reversed for 6 h at 65 °C and the chromatin purified by treatment with RNase A, followed by proteinase K digestion and phenol:chloroform:isoamyl alcohol extraction. For microarray analysis, ≈5,000 embryos and 50 μL antibody were used for each biological replicate. Three biological replicates were performed for anti-Ntl ChIP-chip. ChIP-PCR was performed on ≈300–500 embryos and 3–5 μL antibody/serum were used. Three ChIP-PCR replicates, separate from the ChIP-chip experiments, were performed. ChIP-PCR assays on bud stage and 10 somite stage embryos were performed in parallel. NRS chromatin immunoprecipitations showed no significant enrichment over input chromatin on genomic microarrays (performed once) or in promoter-specific qPCR reactions.

Expression Constructs. T-domain expression constructs were pSP64T Ntl, CS2+ Eomesodermin, CS2+ Spadetail/Tbx16 and CS2+ Tbx6-myc tag, CS2+ Bra. pSP64T, CS2+Ntl-VP16, CS2+ Ntl-EnR. pSP64T Ntl was linearized with SalI, all CS2+ constructs were linearized with NotI. Linearized constructs were transcribed with SP6 polymerase to generate capped mRNA according to standard protocols.

In Vitro Translation, SDS/PAGE and Western Blot Analyses. Capped mRNA for each T-domain protein was used to synthesize ^{35}S-labeled protein in reticulocyte lysate according to manufacturer's instructions (Promega). Reticulocyte lysates were subjected to SDS/PAGE on a 10% gel then blotted onto nylon membrane (GE Healthcare) according to standard protocols. Membranes were blocked overnight in 5% skimmed milk powder in phosphate buffer saline/0.1% Tween 20 (PBST), washed in PBST, incubated for 1 h in a 1:2500 dilution of anti-Ntl antibody in PBST, washed extensively in PBST, incubated for 40 min in a 1:10,000 dilution of HRP-conjugated anti-Rabbit antibody (Pierce) in PBST, and washed extensively in PBST before detection with SuperSignal West Pico Rabbit IgG detection kit (Pierce). For reticulocyte lysate analysis, once protein detection was complete the nylon blot was air dried and exposed to X-ray film to detect ^{35}S-radiolabelled input. (Fig. S2*A*)

ChIP-Chip DNA Amplification, Labeling and Hybridization. For ChIP-chip, purified DNA from anti-Ntl chromatin immunoprecipitation was blunted using T4 polymerase and ligated to linker, as was whole input sonicated chromatin from the same experiment that had not been immunoprecipitated. DNA was then amplified using a 2-stage PCR amplification protocol (1). Amplified DNA was labeled and purified using Bioprime Array CGH random prime labeling and purification kit (Invitrogen). The input chromatin sample was labeled with Cy3; anti-Ntl sample was labeled with Cy5. Labeled DNA for each channel was combined and hybridized to arrays in Agilent hybridization chambers for 40 h at 40 °C. Arrays were then washed and scanned.

Anti-Ntl Antibody. Anti-Ntl antibody was obtained from Stefan Schulte-Merker (2). Previous studies have shown that this antibody is specific to Ntl (2). Recently, a second ortholog of *brachyury* in the zebrafish genome has been described (3), so we tested whether the anti-Ntl antibody would recognize this protein and found that it does not (Fig. S2*A*).

Twelve-Kilobase Genomic Microarrays. We previously described the ChIP-chip technique in zebrafish and the design of promoter microarrays that cover a 2-kb region around the transcription start sites (TSSs) of ≈11,000 zebrafish genes (1). To capture additional regulatory information farther from the basal promoter region, we also designed an expanded set of 60-mer probes that cover 9 kb upstream and 3 kb downstream of the TSS. The 60-mers were chosen so that promoter regions contained approximately 1 probe every 250 bp with a maximum distance between probes for each promoter region set at 600 bp. We also incorporated several sets of control probes, both positive and negative. On each array there are 753 probes designed against seven "gene desert" regions. In addition, because our main motivation for making these microarrays is to identify mesodermal genes regulated by T-box factors we included probes designed to flank the promoter of 7 genes expressed in mesoderm during gastrulation, which preliminary analysis or literature suggested may be targets of Ntl or Spt (*wnt11*, *flh*, *vent*, *msgn1*, *myod*, *fgf8*, and *pcdh8*) and these probes were arrayed on each slide. We also included in this design 502 intensity-control probes. Finally, there are 2,256 controls added by Agilent. The final design contained 378,002 probes, including 365,537 experimental probes divided between 9 microarray slides: slides 1–8 contained 40,616 experimental probes and slide 9 contained 40,609. Further information on design and manufacture of the microarrays can be found in ref. 1 and on the Agilent Technologies Website (www.agilent.com).

Probes were mapped to the Zv6 genome assembly and annotated as associated with a gene if they fell within −9 kb and +3 kb of the TSS; the Zv6 probe locations and annotations are used in all of the analysis described in this study. We excluded probes from the gene list if they were found to map to more than 1 position in the genome unless manual inspection of the mapping revealed this was because of artificial duplication of the genome region because of difficulties in assembling the zebrafish genome. A list of enriched probes and associated genes is given in Table S1.

Array platform files and raw data have been submitted to GEO (www.ncbi.nlm.nih.gov/geo/) with the following accession number: GSE12331.

ChIP-Chip Data Analysis. To identify enriched probes, data were analyzed as previously described (1, 4). This analysis consisted of normalizing the arrays to a set of negative control spots (gene desert controls) and intensity-control spots (described above), which allowed us to normalize across all slides. To correct for different amounts of genomic and immunoprecipitated DNA hybridized to a microarray, the median intensity value of the IP-enriched DNA channel was divided by the median of the input genomic DNA channel, and this normalization factor was applied to each intensity in the input genomic DNA channel. The log of the ratio of intensity in the IP-enriched channel versus the input genomic DNA channel was then calculated for each probe and a whole chip error model (5) was used to calculate confidence values for each spot on each microarray. This error model converts the intensity information to X scores which are assumed to be normally distributed, allowing for the calculation of a *P*-value for the enrichment ratio seen at each feature. To identify bound probes we selected an X score cutoff that would give 5% false positives, assuming a normal distribution of nonbound probes on each slide. Any probe on a slide that returned an X score of greater than or equal to twice the standard deviation of that slide was included. Next we took neighboring probes into account and calculated an average X score for a probe and its immediate neighbor on either side and then calculated a *P*-value for each group of 3 neighboring probes (probe set *P*-values). We required that multiple probes in the probe set provide evidence of binding, so that if the probe set *P*-value was less than or equal to 0.001 the central probe of that set was marked as bound. Probes that satisfied these criteria were mapped to Zv6 and a gene was then annotated as bound by Ntl, and included on our list of "bound" genes, if the enriched probe fell within −9 kb or +3 kb of the TSS of that gene (see also above). It should be noted that of the probes that satisfy all these criteria, 52 probes fall between 32 pairs of genes that lie head to head on the genome (64 genes total). Consequently for these targets, the annotation potentially introduces 50% false positives.

All microarray analysis will lead to identification of some false positives and the exclusion of true positives, so we chose cutoff values to limit false negatives while including the majority of known Ntl targets. At the time of hybridization and data analysis, there were no published targets of Ntl, but our own preliminary work using ChIP-PCR on candidate genes suggested that *flh*, *wnt11*, *fgf8*, *myod*, *msgn1*, *and vent* were targets. The cutoff we used does not identify *fgf8* as a target but does identify our other preliminary targets. At more stringent cutoffs *myod* and *vent* are not identified as targets, while less stringent cutoffs result in an increase in promoters that cannot be verified by ChIP-PCR.

Validation of Microarrays and Target Promoters. To assess the reproducibility of our hybridizations, we included control probes on each slide designed against genomic regions around 7 mesodermal genes (*flh*, *wnt11*, *vent*, *msgn1*, *myod*, *fgf8*, and *pcdh8*) that our preliminary data had suggested may represent direct targets of Ntl or other T-domain factors. Fig. S2*B* shows that the array hybridizations were successful and that the data are consistent across all 9 slides. Of those control mesodermal promoter regions, *flh*, *wnt11*, *vent*, *myod*, and *msgn* are called bound by Ntl in our data analysis using the cutoff values described above. During the preparation of this manuscript, other direct targets of Ntl were identified by 2 other groups (3, 16) and we were reassured to see that all these targets of Ntl (*wnt8*, *dld*, and *tbx6*), although not *wnt3a*, which is not represented on our microarrays, are identified by our analysis, further validating our approach and indicating our data are of high quality. Similarly, functional or sequence orthologs of known *Xenopus* targets such as *wnt11* (*Xwnt11*), *vox* (*Xom*), and *fgf24* (*fgf4*) are also identified as targets in our study.

False Negative and Positive Rates. It is problematic to estimate false negative and false positive rates because this relies on prior knowledge of all genes that are bound or not bound by Ntl in the gastrula stage embryo. Previous reports using similar arrays in yeast that has well-characterized transcription factor binding suggest false positive rates are under 10% and false negative rates lie between 20 and 30% (e.g., ref. 6). We performed ChIP followed by qPCR to verify that the regions identified by ChIP-chip are also identified by conventional ChIP-PCR and that the cutoff values used in the microarray were set at a reasonable level. Primers were designed to 27 genomic regions called bound and 13 genomic regions called not bound in our data analysis and used in a qPCR assay to assess enrichment (Fig. S2*C*). Twenty-three of the "bound" regions were found to be more enriched than the "not bound" regions, and 10/13 not bound regions were less enriched than the "bound promoters."

qPCR. Promoter-specific qPCR was carried out on a Lightcycler 480 using SYBR Green 1 Master kit (Roche) according to manufacturer's instructions, using 55 °C annealing temperature. For each amplicon, a dilution series of wild-type zebrafish genomic DNA from 1 ng to 1 pg was used for a standard curve. For each amplicon, relative enrichment was calculated as follows: [(anti-Ntl sample–no antibody sample)/whole cell extract input sample]. Because negative regions vary in their relative enrichment, we did not normalize to a negative region. Rather, for comparison, negative regions are shown in addition to enriched regions for each experiment. Amplified genomic regions 1–41 with primer sequences can be found in Table S7.

GO Term Analysis. Enrichment analysis was performed using GOToolBox (ref. 7; http://burgundy.cmmt.ubc.ca/GoToolBox/). Searches were done for biological process and molecular function; enrichment analysis included hypergeometric testing and Bonferroni correction of *P*-values.

For biological process, categories with fold enrichment >2 and *P*-values of $<10^{-4}$ were:

pattern specification process (GO:0007389; $P < 1.16^{-21}$; 13.6-fold enriched)

cell fate commitment (GO:0045165; $P < 8.80^{-05}$; 12.1-fold enriched)

regulation of developmental process (GO:0050793; $P < 5.35^{-05}$; 10.3-fold enriched)

embryonic development (GO:0009790; $P < 9.15^{-18}$; 9.2-fold enriched)

multicellular organismal development (GO:0007275, $P < 3.10^{-20}$; 5.0-fold enriched)

cellular developmental process (GO:0048869; $P < 1.27^{-08}$; 4.9-fold enriched)

anatomical structure morphogenesis (GO:0009653; $P < 3.67^{-08}$; 4.8-fold enriched)

anatomical structure development (GO:0048856; $P < 7.08^{-15}$; 4.7-fold enriched)

regulation of metabolic process (GO:0019222; $P < 5.23^{-14}$; 3.8-fold enriched)

regulation of cellular process (GO:0050794; $P < 6.00^{-19}$; 3.6-fold enriched)

regulation of biological process (GO:0050789; $P < 1.08^{-17}$; 3.3-fold enriched).

For molecular function, categories with fold enrichment >2 and *P*-values of $<10^{-4}$ were:

transcription factor activity (GO:0003700; $P < 7.30^{-14}$; 4.8-fold enriched)

transcription regulator activity (GO:0030528; $P < 4.97^{-07}$; 3.2-fold enriched)

nucleic acid binding (GO:0003676; $P < 1.02^{-13}$; 2.8-fold enriched).

Motif Finding. Promoter regions were analyzed by nestedMICA (8) and Trawler (9). For each promoter region with at least 1 positive probe, the most significant probe was chosen as a region on which to center a 500-bp window. The resulting sequences were analyzed by nestedMICA (8) with the following parameters: −targetLength = 8, 10, or 12, −numMotifs = 3,5 or 10, −revComp flag set. The results were not affected by changes to the target length or the number of motifs. The background model for nestedMICA was created by using negative matched regions.

Negative matched regions were created from CRMs that did not contain an enriched probe (using a low stringency cutoff of probe P-value ≤ 0.01 and a probe set P-value ≤ 0.01). Negative regions were matched exactly to the CG content and the repetitive sequence percentage of the positive sequence regions. Additionally, the negative regions were taken from a region between 6,000 bp upstream and 1,500 bp downstream of the TSS to avoid regions at the edges of the arrayed promoters.

Motif Mapping. Putative binding sites for Ntl were identified by generating all possible sequences corresponding to the position weight matrix shown in Table S5 with a bit score (10) greater than 0 using an in-house written perl script. These putative binding sequences were then mapped, allowing no mismatches, to the CRM sequences using Exonerate (11). The position weight matrix (PWM) in Table S5 is based on the mouse T-BOX motif from JASPAR; however, a small pseudocount is added to all 0 values allowing the occurrence of sequences observed in the motif-finding exercise, which are precluded by the 0 values in the JASPAR PWM. Unless otherwise indicated, mappings were rejected unless the nucleotide at position 2 was T and the nucleotides at positions 5, 6, and 7 were C, A, and C, respectively. The remaining mappings were termed "constrained." The same method was used for other JASPAR vertebrate PWMs, however pseudocounts were not added and no constraints were applied.

Conservation Analysis. Conservation scores were downloaded from the UCSC Genome Browser (hgdownload.cse.ucsc.edu/goldenPath/danRer4/phastCons7wayScores). These scores are calculated by the PhastCons algorithm (CITE) based on a 7-way alignment of the following species' genome assemblies: zebrafish (Mar. 2006), Fugu (Aug. 2002), Tetraodon (Feb. 2004), frog (Aug. 2005), opossum (Jan. 2006), mouse (Feb. 2006), human (Mar. 2006, hg18). PhastCon scores represent the posterior probability that the column in the multiple alignment was generated from a conserved state and can be interpreted as the probability that the base is in a conserved element (12). For maximum sensitivity, we used conservation scores greater than 0.1 for all analysis (i.e., bases with a probability of at least 10% of being in a conserved element).

Conservation (i.e., phastCons) values for the regions of all of the CRMs that met our conservation score threshold and were not identified as containing the target motif were designated as the average background conservation of the CRMs. Using a t-test, the conservation scores from the identified motif location were compared to the background CRM conservation. Motif P-values ≥ 0.05 were considered significant. Note that this technique may underestimate the evolutionary constraint in the motif because the CRM outside of the identified motif is likely to contain both evolutionarily constrained regions associated with other binding sites and regions that are not subject to evolutionary constraint.

We also used 2 other methods to look for conservation of the T-binding motif in other species. First, we used TRAWLER (9) to search for conservation in other fish species. This method uses pairwise alignments between promoter sequences and their orthologs in other species, but this approach did not show significant conservation in our promoter sequences. Second, we looked for clustering of the T-binding motifs in our promoter sequences using an approach similar to ref. 16. This confirmed the observation that cdx4 is conserved in its upstream region, together with regions upstream of *drl, kpna2, irx3a*, and *fibpl*, of which all but *drl* also appear on our PhastCons list of conserved CRMs.

In situ Hybridization. Whole mount in situ hybridization on zebrafish embryos was carried out as described (13). Antisense probes were generated from vectors, generously provided by numerous colleagues, by linearization with the appropriate restriction enzyme and transcription with the appropriate RNA polymerase. Details of these can be obtained on request.

Promoter Constructs and Luciferase Assays. Genomic regions were amplified from genomic DNA prepared from AB strain zebrafish by PCR using proof-reading KOD XL polymerase (Novagen), cloned into pCR-BluntII-TOPO vector (Invitrogen), and then subcloned into pGL3-Basic (Flh:2067) or pGL3-Promoter (Flh TBS1 + 2) (Promega).

One-cell embryos were injected with 40 pg luciferase construct and 0.75 pg pCS2 + Renilla as an injection control, with or without wild-type Ntl mRNA or Ntl-VP16 mRNA. For wild-type Ntl mRNA 150 pg, 375 pg, or 750 pg were used; for Ntl-VP16, 150 pg mRNA was used. However 750 pg of wild-type Ntl mRNA resulted in high levels of lethality. Thirty to 50 embryos were collected at 75–85% epiboly and homogenized in 10 μL/embryo of 1X Passive lysis buffer (Promega). Samples were then diluted 10-fold and quantified using the Dual Luciferase Assay kit (Promega). Each experiment was performed a minimum of 3 times. All data are reported as the mean fold change in luciferase activity compared to the condition where no mRNA was injected and reported with standard error of the mean. Differences in the luciferase activity for different constructs were compared by a Mann–Whitney U-test. A P-value ≤ 0.05 was considered significant.

Morpholino Injections. One-cell embryos were injected with 0.25 pmol of Ntl morpholino (ref. 14; Genetools).

GRN. The GRN was drawn using Biotapestry software (15).

1. Wardle FC, *et al.* (2006) Zebrafish promoter microarrays identify actively transcribed embryonic genes. *Genome Biol* 7: R71.
2. Schulte-Merker S, Ho RK, Herrmann BG, Nusslein-Volhard C (1992) *Development* 116:1021–1032.
3. Martin BL, Kimelman D (2008) Regulation of canonical Wnt signaling by Brachyury is essential for posterior mesoderm formation. *Dev Cell* 15:121–133.
4. Boyer LA, *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122:947–956.
5. Hughes TR, *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102:109–126.
6. Lee TI, *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* 298:799–804.
7. Martin D, *et al.* (2004) GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biol* 5:R101.
8. Down TA, Hubbard TJ (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res* 33:1445–1453.
9. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods* 4:563–565.
10. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) in *Biological Sequence Analysis.* (Cambridge Univ Press, Cambridge, UK).
11. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
12. Siepel A, *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
13. Nusslien-Volhard C, Dahm R (2002) *Zebrafish* (Oxford Univ Press, Oxford, UK).
14. Feldman B, Stemple DL (2001) Morpholino phenocopies of sqt, oep, and ntl mutations. *Genesis* 3:175–177.
15. Longabaugh WJ, Davidson EH, Bolouri H (2005) Computational representation of developmental genetic regulatory networks. *Dev Biol* 283:1–16.
16. Garnrtt AT, *et al.* (2009) Identification of direct T-box target genes in the developing zebrafish mesoderm. *Development* 136:749–760.

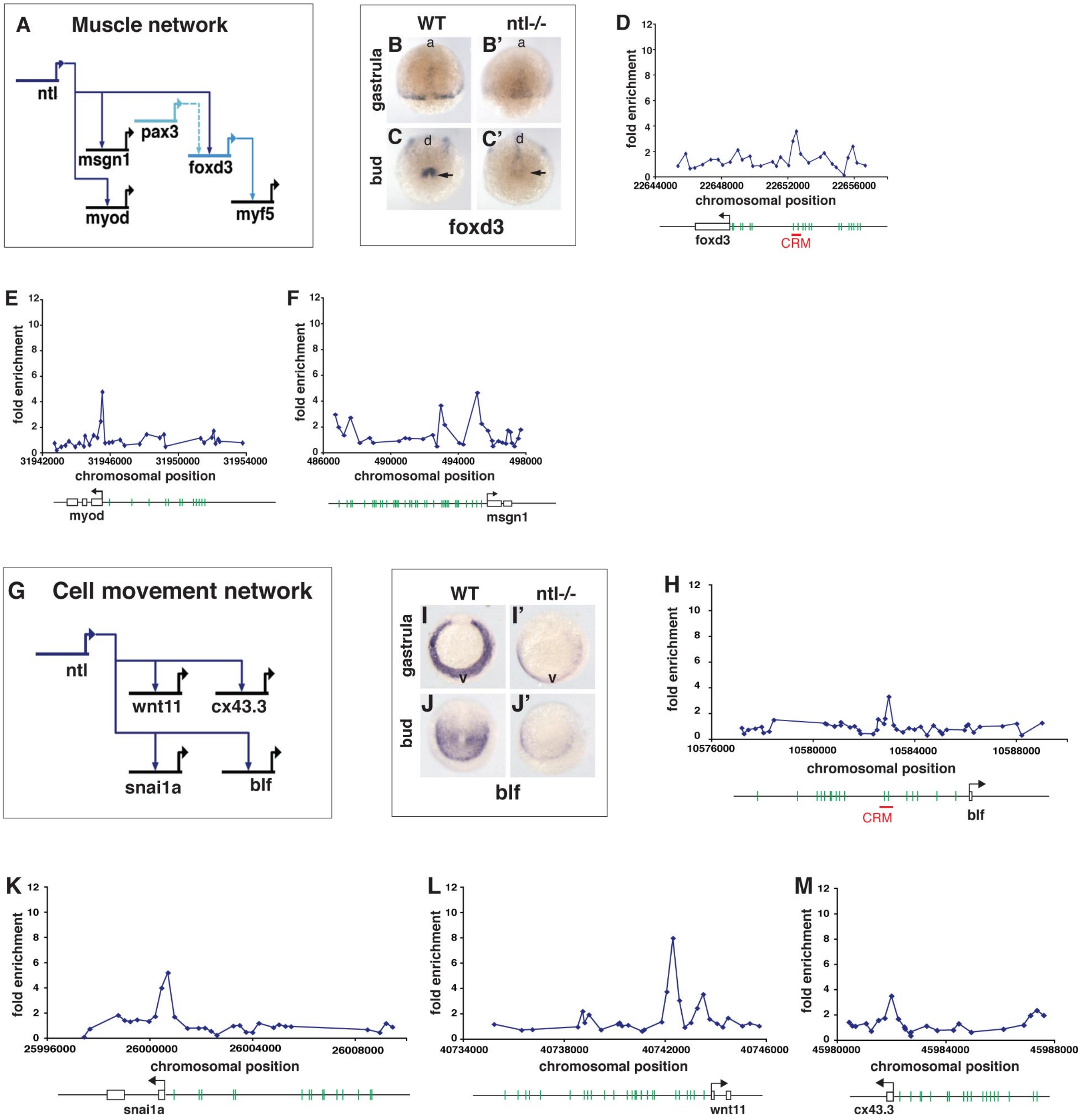

Fig. S1. Ntl binds and regulates expression of genes involved in muscle specification and morphogenetic movements. (*A*) GRN showing Ntl binding and regulation of factors involved in specifying muscle cell fate. Solid lines indicate binding of target promoter *and* genetic regulation. Dashed line indicates target binding only *or* genetic regulation without proven target binding. (*B–C*′) In situ hybridization of *ntl* mutants compared to wild types at gastrula and bud stages shows downregulation of *foxd3* (arrows). (*D–F*) Ntl binding in genomic regions around *foxd3*, *myod*, and *msgn1*. Plots show ChIP-enrichment ratios for microarray probes in the genomic regions shown. Chromosomal position, TSS, intron-exon structure, putative upstream T-binding sites with constrained character (green lines; see text), and conserved CRM (red bar; see text) are shown below the graphs. (*G*) GRN showing Ntl binding and regulation of genes involved in morphogenetic movements. (*I–J*′) In situ hybridization shows downregulation of *blf* in *ntl* mutant compared to wild-type embryos at 70% epiboly and bud stage. (*H–M*) Ntl binding in genomic regions around *blf, snai1a*, *wnt11*, and *cx43,3*. a, anterior; d, dorsal; v, ventral

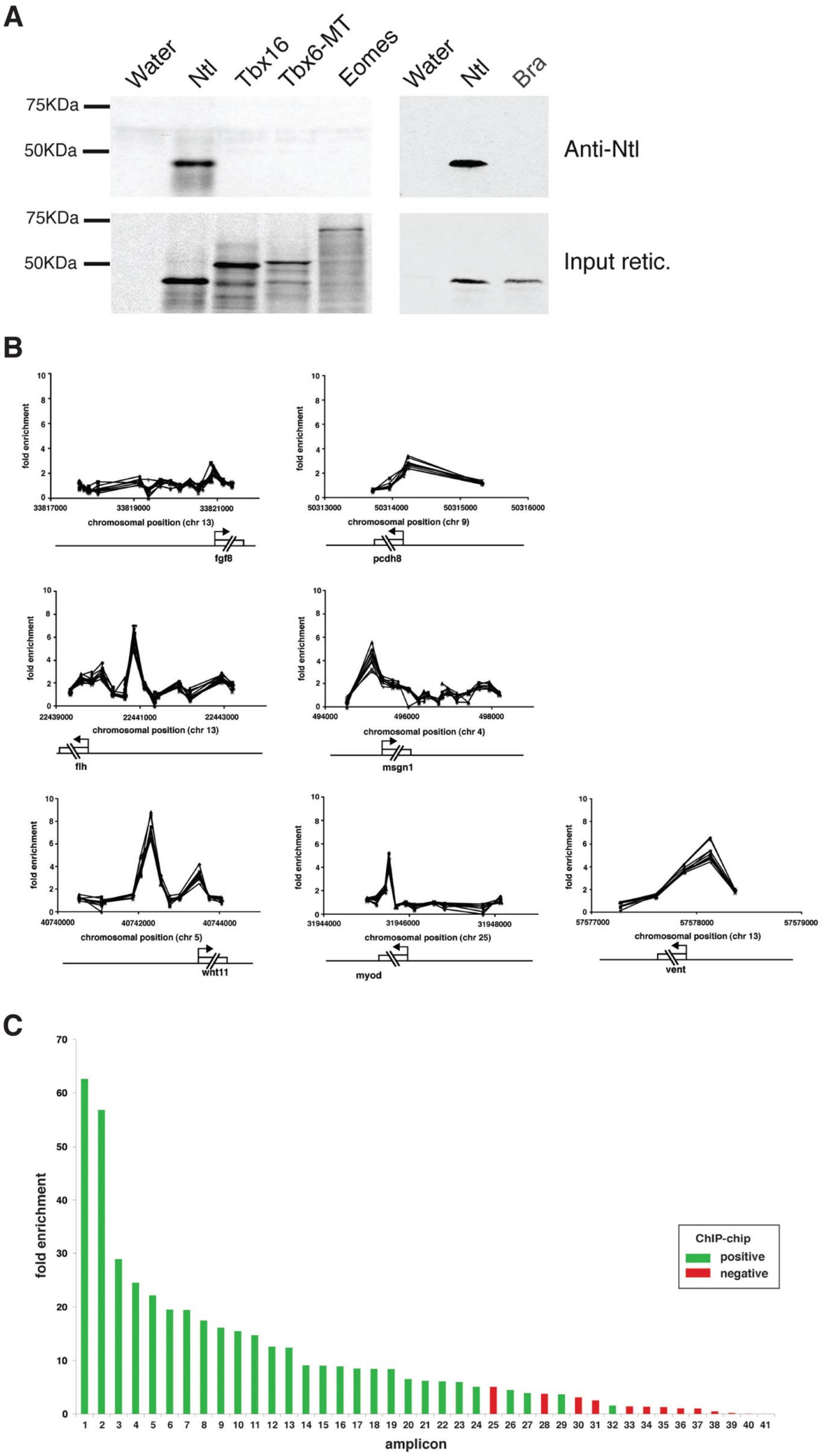

Fig. S2. Specificity of anti-Ntl antibody and validation of microarrays. (*A*) Anti-Ntl antibody specifically recognizes Ntl protein. Western blots with anti-Ntl antibody on different T-box proteins translated in vitro in reticulocyte lysates. Ntl protein is detected by the antibody but Tbx16, Tbx6-myc tag, Eomes, and Bra are not. Input translation products (^{35}S-labeled) are shown below the Western blots. (*B*) Positive control regions show consistent enrichment across all 9 slides in microarray hybridizations. Plots show median ChIP-enrichment ratio for microarray probes on each of 9 slides in the genomic region shown below the plot. Chromosomal position and transcription start site of each gene are shown below the plot. (*C*) The majority of genomic regions called ‘‘bound’’ (positive) by ChIP-chip are confirmed as enriched by ChIP-PCR compared to genomic regions called ‘‘not bound’’ (negative). Amplified genomic regions 1–41 with primer sequences can be found in Table S7.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6
Table S7